Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Mateja Volčanšek

Leksikalna analiza razpoloženja za slovenska besedila

diplomsko delo

visokošolski strokovni študijski program prve stopnje računalništvo in informatika

Mentor: izr. prof. dr. Janez Demšar

Ljubljana 2015

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Analiza razpoloženja (sentiment analysis) je področje računalniške znanosti, konkretno tekstovnega rudarjenja, ki se ukvarja z odkrivanjem mnenj in čustev, ki ga odražajo besedila. Področje se je posebej razvilo v zadnjem desetletju, saj omogoča avtomatsko obdelavo množice besedil na spletu, predvsem na različnih forumih in v komentarjih novic.

Kot v večini tekstovnega rudarjenja so tudi orodja za analizo čustev omejena predvsem na angleški jezik. V diplomski nalogi raziščite, kako si lahko z njimi pomagamo pri analizi slovenskih besedil, bodisi s prevajanjem slovarja čustveno opredeljenih besed v slovenščino bodisi s prevajanjem slovenskih besedil v angleščino in analizo čustev v angleščini. Za potrebe naloge pripravite tudi primeren slovar besed in korpus označenih slovenskih besedil.

Izjava o avtorstvu diplomskega dela

Spodaj podpisana Mateja Volčanšek sem avtorica diplomskega dela z naslovom:

Leksikalna analiza razpoloženja za slovenska besedila

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Janeza Demšarja,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 29. januarja 2015 Podpis avtorja:

Zahvaljujem se mentorjema za strokovno pomoč, posebej dr. Petri Kralj Novak za idejo, vodenje in podporo pri izdelavi diplome ter svojim staršem, ki so mi študij omogočili in me vseskozi podpirali.

Kazalo

Seznam uporabljenih kratic

CA (classification accuracy) - Klasifikacijska točnost
DBMS (database management system) - Sistem za upravljanje podatkovnih baz
SVM (support vector machine) - Metoda podpornih vektorjev
HTML (Hypertext Markup Language) - Označevalni jezik, ki je bil razvit za prikazovanje podatkov, s poudarkom na izgledu podatkov
CSS (Cascading Style Sheets) - Kaskadne stilske podloge, preprost mehanizen za dodajanje stilov, barv, ozadij spletnim stranem
PHP (Hypertext Preprocessor) - Odprto kodni skriptni programski jezik, ki je splošno namenski, a posebej primeren za razvoj dinamičnih spletnih strani
WWW (World Wide Web) - Informacijski sistem notranje povezanih hipertekstnih (nadbesedilnih) dokumentov
XML (Extensible Markup Language) - Označevalni jezik, ki je bil razvit za opisovanje podatkov in prenašanje teh med omrežji
API (Application programming interface) - Skupek protokolov in orodij za gradnjo programskih aplikacij

Povzetek

Diplomsko delo vsebuje opis izdelave slovenskega slovarja za zaznavo subjektivnih elementov v besedilu, ki se uporablja v leksikalnih metodah za avtomatsko analizo razpoloženja. Zanima nas, kako učinkovita je uporaba slovarja, ki ga prevedemo iz že obstoječega angleškega slovarja, za ugotavljanje razpoloženja v slovenskih besedilih. Učinkovitost preverimo na korpusu 5000 ročno označenih besedil, ki smo jih zajeli iz glavnih slovenskih spletnih portalov novic. Rezultate primerjamo z alternativno metodo za neangleška besedila: korpus prevedemo v angleščino in nato naredimo analizo razpoloženja.

V diplomskem delu je najprej predstavljen pojem analize razpoloženja, njena uporabnost in razlogi za razširjenost. V nadaljevanju se osredotočimo na tehnike analize, predstavimo metode, s katerimi si lahko pomagamo, in poudarimo pomembnost leksikalnih virov ter pomanjkljivost prosto dostopnih virov v slovenskem jeziku.

Glavni del diplomske naloge predstavlja opis izdelave slovenskega slovarja za zaznavo elementov subjektivnosti s pomočjo prevajalskih orodij in že obstoječega slovarja v drugem jeziku ter opis analize razpoloženja. V okviru dela sta nastala slovar slovenskih besed in korpus slovenskih novic, ki je ročno označen glede na polarnost besedila (pozitivno (angl. positive), negativno (angl. negative), nevtralno (angl. neutral)).

Ključne besede: tekstovno rudarjenje, analiza razpoloženja, leksikalna metoda, mnenje, subjektivnost.

Abstract

The goal of this thesis is to create a sentiment dictionary for the Slovenian language which can be used in lexical methods for automatic sentiment analysis. We start from a sentiment dictionary for the English language, translate it semi-automatically to Slovenian and curate its content. We test the performance of using the translated dictionary for automated lexical sentiment analysis on a corpus of 5000 manually annotated Slovenian news articles gathered from the main Slovenian news portals. The results of the analysis are compared with the results of an alternative method, where, instead of translating the sentiment dictionary, the documents are translated to English and lexical sentiment analysis is performed.

This thesis is organized as follows. First, the concept and motivation for automated sentiment analysis are introduced. Next, the techniques for sentiment analysis are outlined, stressing the importance of sentiment dictionaries in automated sentiment analysis. The main part of the thesis is Chapter 4, in which the process of creating the Slovenian sentiment dictionary is described and explained in detail. Furthermore, the manual article annotation process is described and the experimental evaluation of the two alternative methods is performed.

Within the practical part of this thesis, a Slovenian sentiment dictionary and a manually annotated corpus of 5000 Slovenian news articles were created.

Keywords: text mining, sentiment analysis, lexicon-based method, opinion, subjectivity.

Poglavje 1

Uvod

Čustva, razpoloženje, mnenja, pogledi posameznikov imajo pomembno vlogo v medčloveških odnosih, tako v vsakdanjem kot v poslovnem življenju. Pri pogovoru te atribute razkriva višina in barva glasu, pri srečanjih pa jih lahko spremljamo tudi z opazovanjem obrazne in telesne mimike. Pri besedilih je odkrivanje in izražanje čustev težje, saj je izražanje omejeno na besede. Avtorji besedil lahko vplivajo na razpoloženje s previdno izbiro besed, pridevnikov, glagolov, z oblikovanjem stavkom, pisanja v tretji ali prvi osebi [25].

Z razvojem spleta, družabnih omrežij, spletnih trgovin, forumov, spletnih dnevnikov, knjižnjic se je komunikacija med posamezniki preselila na svetovni splet. Dandanes je uporaba spleta tako poenostavljena in dostopna, da lahko preko spleta komunicirajo vsi. V trenutku lahko vsakdo pusti jezno sporočilo prodajalcem produkta, ki se mu je pokvaril, graja trenutne odločitve politikov z uporabo družabnega omrežja kot je Twitter ali proslavi zmago svojih najljubših športnikov z zapisom na spletnem dnevniku. Da sporoči svoje razpoloženje, mnenje in čustva, se mu ni treba premakniti iz hiše ali sploh srečati ljudi, katerim so ta sporočila namenjena. Zaradi udobnosti, enostavnosti, časovne učinkovitosti in do neke mere tudi anonimnosti, ki jo splet ponuja, je količina javno dostopnih besedil zelo narasla in se dnevno povečuje. Posebnost spletnih besedil je, da nastajajo vsako sekundo, so nestrukturirana, zapisana v naravnem jeziku, predvsem pa vsebujejo veliko povratnih in-

formacij o kakovosti izdelkov, storitev, naklonjenosti do politikov, športnikov, znanih oseb, podjetij in odražajo mnenje o trenutnih političnih in gospodarskih razmerah. Včasih se je javno mnenje lahko raziskovalo le preko anket (osebno, preko telefona ali pošte), sedaj pa so se odprle nove možnosti spremljanja mnenja, in to takoj, ko avtor objavi svoje mnenje [25].

Kljub velikemu napredku in lahki dostopnosti informacij preko medomrežja, dejstvo, da je podatkov ogromno, so neurejeni in velikokrat tudi nekoristni, ostaja nespremenjeno. Posameznik težko najde koristne informacije med naraščajočo maso podatkov, ki se vsakodnevno steka na medomrežje. Prav zaradi potrebe po avtomatizaciji iskanja in urejanja mnenj se je razvila analiza razpoloženja.

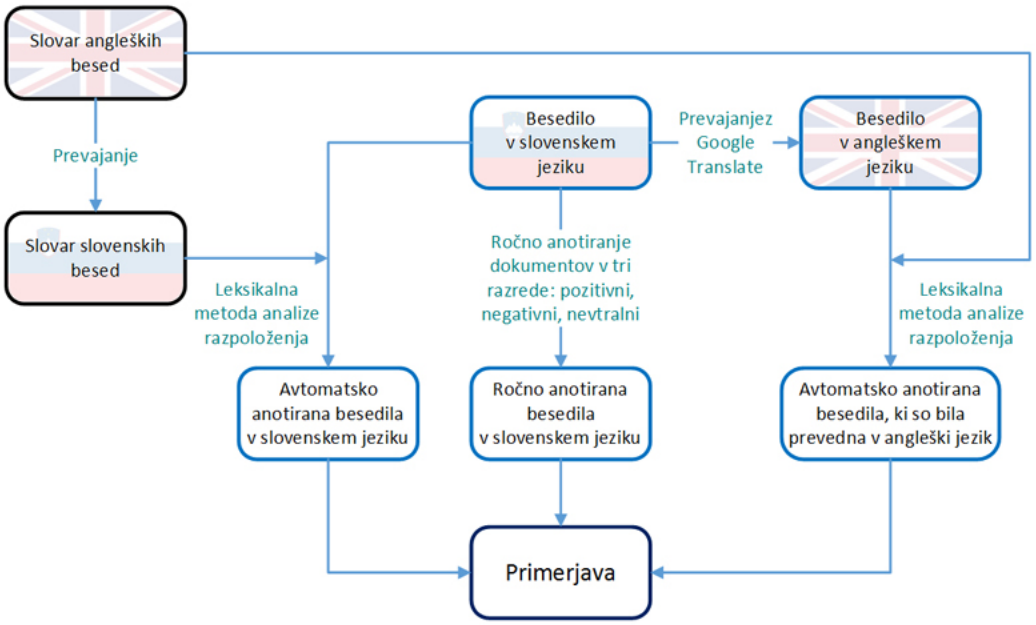
Analiza razpoloženja ali rudarjenje mnenja (angl. sentiment analysis) je področje znanosti, ki preučuje človekova mnenja, odnos, čustva, razpoloženje oziroma pogled na določene entitete ter njihove atribute. Pri tem se analiza opira na obdelavo naravnega jezika, analizo besedil in računalniško lingvistiko. Pojem entiteta lahko zajema produkte, storitve, organizacije, posameznike, dogodke, teme, besedila [13].

1.1 Motivacija

Zaradi razširjenosti angleškega jezika je večina raziskav in virov na področju rudarjenja mnenja narejena za angleški jezik. To vključuje korpuse, slovarje in sisteme za avtomatsko detekcijo mnenja. Zaradi pomanjkanja ali nedostopnosti jezikovnih orodij za analizo razpoloženja v slovenskem jeziku želimo prispevati k razvoju teh in izdelati slovar ter korpus v slovenskem jeziku. Slovar tudi preizkusimo z leksikalno metodo in skušamo rezultate uporabiti za njegovo izboljšanje. Ker je izdelava slovarjev zelo časovno potratna in zahtevna, si pri izdelavi pomagamo s slovarjem v angleškem jeziku. Ideja raziskave je, da je enostavneje enkrat prevesti angleški slovar in ga uporabljati na slovenskih besedilih kot pa kot pa vedno znova prevajati korpuse in uporabljati angleški slovar. V nalogi želimo tudi preveriti, kako dobro

se da uporabiti leksikalno metodo in slovar, narejen v tujem jeziku, v tem primeru angleškem, na slovenskih besedilih.

V praktičnem delu diplomskega dela bomo pripravili slovenski slovar, izdelali spletno aplikacijo za ročno ocenjevanje besedil, s katero bomo nato ocenili zajeta besedila. Na teh besedilih bomo nato preverili uspešnost avtomatske metode s slovenskim slovarjem in pa uspešnost avtomatske metode z angleškim slovarjem. Rezultate bomo analizirali in jih skušali uporabiti za izboljšavo slovenskega slovarja. Slika 1.1 prikazuje shemo naše analize: Besedilo v slovenskem jeziku (a) ročno označimo, (b) prevedemo v angleški jezik in naredimo avtomatsko leksikalno analizo razpoloženja in (c) uporabimo v slovenščino preveden slovar subjektivnih besed in naredimo leksikalno analizo razpoloženja v slovenskem jeziku. Rezultate vseh treh pristopov primerjamo.

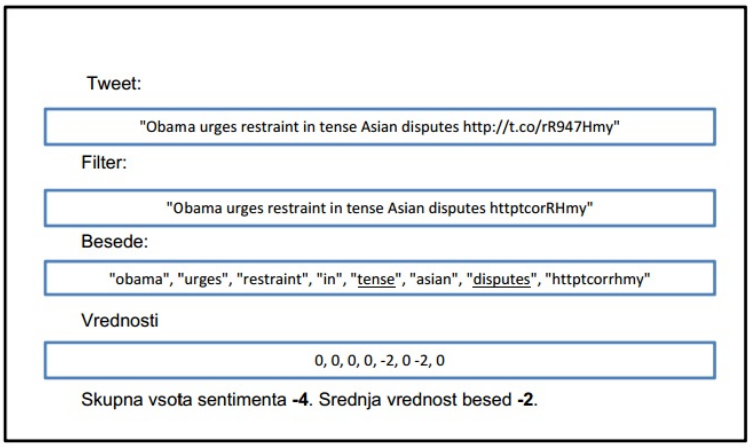


Slika 1.1: Shema raziskave.

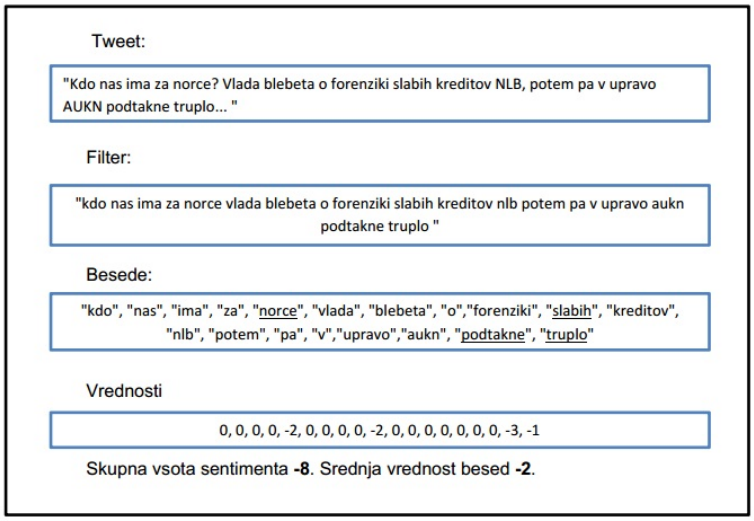
1.2 Obstoječi pristopi analize razpoloženja za slovenski jezik

Prosto dostopnih pristopov za slovenski jezik na področju analize razpoloženja zelo malo. Našli smo le dve obsežnejši deli. Delo, ki opisuje izdelavo slovarja za slovenski jezik na podoben način in primerjavo med angleškim in slovenskim jezikom, je magistrsko delo Roka Martinca z naslovom Merjenje sentimenta na družabnem omrežju Twitter: izdelava orodja in evalvacija [15]. V svojem delu opiše, kako je naredil analizo razpoloženja preko družabnega omrežja Twitter. Izbiro Twitterja utemelji z dejstvom, da je za takšno analizo najbolj primeren zaradi enostavnega API-ja in zaradi veliko uporabnikov, ki svojo misel oz. počutje spravijo v jedrnate 140 znakovne stavke. Za potrebe dela je sprogramiral algoritem v programskem jeziku R in uporabil paket twitteR. Na Sliki 1.2 in Sliki 1.3 je prikazano delovanje aplikacije za analizo razpoloženja najprej v angleškem in nato v slovenskem jeziku. Slovar je izdelal s pomočjo seznama AFINN-111, ki je narejen posebej za mikrobloge. AFINN-111 je seznam angleških besed in fraz z njihovo valenco med -5 (negativno) do +5 (pozitivno). Najprej je analiziral tvite v angleščini z angleško različico AFINN-111. Nato pa je tabelo AFIN-111 prevedel v slovenski jezik in analiziral slovenske tvite. Analizo razpoloženja je omejil na nekaj tem - predsedniške volitve, Janez Janša, Gregor Virant, Tina Maze. Njegovi rezultati za angleški jezik so primerljivi s tistimi, ki so že na trgu. Rezultati za slovenski jezik so bili pomanjkljivi zaradi manjše uporabe omrežja Twitter med Slovenci in posledično vhodnih podatkov ter kompleksnosti slovenskega jezika [15].

Drugo delo, ki opisuje pristope analize razpoloženja za slovenski jezik, je diplomsko delo Brine Škoda, z naslovom Rudarjenje razpoloženja na komentarjih rtvslo.si. V tem delu se osredotočijo na metode strojnega učenja in z njimi zgradijo klasifikator za analizo razpoloženja v slovenskem jeziku [11]



Slika 1.2: Prikaz delovanja algoritma Roka Martinca za angleški jezik.



Slika 1.3: Prikaz delovanja algoritma Roka Martinca za slovenski jezik.

Poglavje 2

Tehnike analize razpoloženja

Na analizo razpoloženja vpliva ogromno dejavnikov in je težko ter časovno zahtevno vključiti vse. V nadaljevanju omenimo nekaj najpomembnejših.

Za analizo razpoloženja sta pomembna pojma subjektivnost in objektivnost. Subjektivnost pomeni nekaj, kar je, obstaja odvisno od človekove zavesti, mišljenja, medtem ko objektivnost pomeni nekaj, kar je, obstaja neodvisno od človekove zavesti, mišljenja [20].

Elementi subjektivnosti so gradniki besedila, s katerimi avtor opiše svoje mnenje, čustva do entitete. Med elemente subjektivnosti v našem delu prištevamo besede, ki se nahajajo v slovarju. Poleg teh besed bi k elementom subjektivnosti spadale tudi besedne zveze ter posebni znaki oziroma simboli, ki se na spletu uporabljajo za izražanje čustev. Subjektivnost se navadno eksplicitno izraža z izbiro nedvoumnih besed in besednih zvez, vendar jo je velikokrat težko zaznati na primer, kadar avtor uporabi sarkazem, ironijo ali pa svoje nezadovoljstvo oziroma zadovoljstvo opiše z implicitnim mnenjem.[25]

Eksplicitno mnenje je subjektivna izjava, ki poda navadno ali primerjalno mnenje. Na primer: *"Kruh pekarne Pečjak je zelo dober."* ali *"Kruh pekarne Pečjak je boljši kot kruh iz Grajskih pekarn."* Implicitno mnenje je objektivna izjava, ki implicira navadno ali primerjalno mnenje. Takšna objektivna izjava navadno izraža želeno ali nezaželeno dejstvo. Na primer: *"Baterija telefona*

znamke Nokia zdrži dlje od baterije telefona znamke Samsung." Eksplicitna mnenja lažje najdemo in klasificiramo.[13].

Na področju analize razpoloženja in rudarjenja mnenja še ne obstaja enotna taksonomija razpoloženja in čustev. Eden najpogostejših načinov, s katerim avtorji predstavijo besedila, je semantična usmerjenost ali polarnost besedila. Usmerjenost besedil, ki vzbujajo pozitivne občutke, je pozitivna (označimo jo z vrednostjo 1), besedil, ki vzbujajo negativne občutke, negativna (označimo jo z z vrednostjo -1) in besedil, ki ne vzbujajo posebnih občutkov, nevtralna (označimo jo z 0). Poleg te metode se uporablja še širok spekter bolj zahtevnih metod, med drugim tudi teorija o delitvi čustev, ki besede znotraj pozitivnega in negativnega spektra delijo še naprej glede na čustva (ljubezen, veselje, presenečenje, jeza, žalost, strah).[25] Ker je pojem mnenja zelo širok, se v tem delu opredelimo na pozitivne, negativne ali nevtralne občutke, ki jih avtor izrazi o entiteti, torej na polarnost besedila.

Pri analizi je treba biti pozoren tudi na perspektivo. Na mnenja lahko gledamo iz dveh perspektiv. Avtorjeve (angl. opinion holder), ki izraža mnenje in bralčeve, ki prebere mnenje. Primer: *"Cene nepremičnin so zopet padle, kar je slabo za gospodarstvo."* Očitno je, da avtor članka govori o negativnem vplivu padca cen na gospodarstvo. Ampak bralci lahko na to novico gledajo iz več zornih kotov. Za prodajalce je novica seveda negativna, toda za kupce izredno dobra, torej pozitivna.[13]

Analiza razpoloženja se lahko v splošnem vrši na treh različnih nivojih:

- Nivo dokumenta: Cilj te analize je določiti ali dokument kot celota izraža pozitivno ali negativno usmerjenost. Za primer lahko vzamemo mnenje o produktu, ki ga je kupec oddal na spletni strani. V interesu podjetja je, da ugotovi, ali je kupec s produktom zadovoljen oziroma če je mnenje kot celota pozitivno ali negativno. Ta analiza predvideva, da vsak dokument izraža mnenja le o eni sami entiteti (produktu).

- Nivo stavka: Cilj te analize je določitev pozitivnega, negativnega ali nevtralnega mnenja za vsak stavek. Če je stavek nevtralen, to pomeni, da ne izraža subjektivnosti.

- Nivo vidika: Analiza na nivoju dokumenta in stavka nam ne pove točno, *kaj* je bilo nekomu všeč ali ne všeč. Zato namesto jezikovnih gradnikov gledamo z vidika razpoloženja. Mnenje je sestavljeno iz razpoloženja (pozitivnega ali negativnega) in tarče mnenja. S tem povečamo natančnost, saj lahko za vsako mnenje, natančno povemo, na katero entiteto se nanaša. V primeru *"Čeprav strežba ni najboljša, mi je ta restavracija še vedno zelo všeč."* lahko opazimo, da stavek kot celota izraža mešano mnenje. Vendar, če se osredotočimo na vsak vidik posebej, vidimo, da je negativno mnenje izraženo glede strežbe, pozitivno mnenje pa izraženo glede restavracije. S tem zvišamo natančnost pri analizi razpoloženja, vendar je tak postopek tudi ustrezno bolj zahteven od drugih.[13]

Metode analize razpoloženja delimo na dve kategoriji: leksikalne metode in metode strojnega učenja.

2.1 Leksikalni pristop in viri

Leksikalne metode temeljijo na leksikalnih virih. To so seznami oziroma slovarji, ki vsebujejo subjektivne besede, besedne zveze, fraze in idiome značilne za pozitivno ali negativno čustvo. Na primer: *'dober'*, *'angelski'* in *'boljši'* so pozitivne besede. *'Slab'*, *'reven'*, *'grozen'* izražajo negativnost. Poleg posameznih besed, subjektivnost izražajo tudi fraze in idiomi na primer: *"Sosed je izgubil glavo."* ali *"Iti rakom žvižgat."*. Nekateri seznami uporabljajo tudi uteži, s katerimi poudarijo bolj močno negativno ali pozitivno čustvo.

Večina orodij za analizo razpoloženja se do neke mere zanaša na seznam besed in fraz s pozitivno ali negativno anotacijo, ali pa so empirično povezane s pozitivnimi oziroma negativnimi komentarji. Taki seznami se navadno ne uporabljajo kar takšni, kot so, ampak se jih modificira in prilagodi domenam, na katerih jih kasneje uporabljamo, da lahko dobimo dobre rezultate. Poznamo tri načine, kako pridobiti slovar. Prvi je ročni način, kjer

ljudje naredijo slovar na roke. Ročni način je zelo zahteven in zelo zamuden, sploh če si želimo namenskega slovarja za vsako domeno. Drugi način uporablja semena besed (ang. seed words), katerih subjektivno usmerjenost že poznamo. S pomočjo teh se nato razširi prvotni seznam z drugimi slovarji, kot recimo WordNet. Besedam poišče vse sinonime in antonime, nato pa se izračuna razdalja med dvema izrazoma. Iz tega se nato izračuna stopnja subjektivnosti besede. Negativna plat tega načina je, da ni zgrajen na določeni domeni in ne vsebuje posebnosti in značilnosti katerekoli domene. Tretji način je podoben drugemu, a pri tem se seznam semen besed razširi s pomočjo korpusa dokumentov iz posamezne domene. Klasičen algoritem, ki se uporablja v tem načinu, je da v korpusu s pomočjo lingvističnih konektorjev (and, or, neither-nor, either-or) poiščemo dodatne pridevnike, za katere lahko z gotovostjo določimo subjektivno usmerjenost.[4] Čeprav so besede izredno pomemben gradnik analize razpoloženja, imajo slovarji določene pomanjkljivosti, na katere je pri njihovi uporabi potrebno paziti:

- Pozitivna ali negativna beseda ima lahko nasprotno orientacijo oziroma pomen v različnih kontekstih. Na primer: beseda *oster* ponavadi izraža negativno mnenje kot na primer: *oster prijem*, *ostra vzgoja*, *oster pogled*, vendar če ga uporabimo v stavku *"Nož je oster."* implicira pozitivno mnenje.

- Stavek sicer vsebuje besedo, ki izraža subjektivnost, vendar je kot celota stavek ne izraža. To se kaže na nekaterih, a ne vseh, primerih vprašalnih in pogojnih stavkov: *"Ali je ta pralni stroj dober?"* in *"Če najdem dober pralni stroj, ga bom kupila."* Oba stavka vsebujeta besedo *'dober'*, ki je pozitivna, ampak noben stavek ne izraža subjektivnega mnenja glede specifičnega pralnega stroja.

- Stavki, ki izražajo sarkazem, so problematični za analizo. Primer: *"Nikoli ne pozabim obraza, vendar v tvojem primeru bom z veseljem napravil izjemo."*

- Veliko stavkov, ki na prvi pogled ne vsebujejo besede, ki izraža sub-

jektivnost, lahko implicira mnenja. Po navadi so to stavki, ki izražajo objektivno a (ne)zaželeno lastnost entitete. Na primer: *"Ta pralni stroj porabi veliko vode."*

Za leksikalni pristop potrebujemo neoznačen korpus (zbirka besedil) in pa leksikalni vir oziroma slovar. Vsaka beseda, ki je v slovarju, je primerjana z besedami v neoznačenem korpusu. Če je beseda iz besedila prisotna v slovarju, potem je subjektivna usmerjenost te besede dodana celotni usmerjenosti besedila. Na primer, če najdemo ujemanje z besedo *odlično*, ki je označena v slovarju kot pozitivna, potem se celotna subjektivna vrednost besedila poveča. Če je celotna subjektivna vrednost besedila pozitivna, potem je besedilo v celoti pozitivno, v nasprotnem primeru pa je negativno. Ker je klasifikacija besedila v celoti odvisna od subjektivne vrednosti, ki jo pridobimo na podlagi slovarja, je velik poudarek na raziskavah slovarja in informacij, ki jih le ta vsebuje. Raziskovalci se poslužujejo najrazličnejših tehnik, od slovarja zgrajenega le iz pridevnikov do izdelave semantične orientacije posamezne besede v slovarju.[1]

2.1.1 General Inquirer

General Inquirer [7] je sistem za analizo besedil. Temelji na uporabi leksikalnega vira, ki je sestavljen iz štirih virov. Vsebuje 182 kategorij besed. Vsaka kategorija je seznam besed, označen s pomenom, besedno vrsto in kategorijami, v katere še spadajo, saj lahko ena beseda spada v več kategorij. Kategoriji Positiv (1914 pozitivno orientiranih besed) in kategoriji Negativ (2290 negativno orientiranih besed) sta najobširnejši in navadno največkrat osnova za izdelavo slovarjev ali služita kot referenca za primerjavo pri avtomatski gradnji leksikalnih virov.[25] Ti dve kategoriji smo v našem delu vzeli kot osnovo za izdelavo slovarja v slovenskem jeziku.

2.1.2 WordNet-Affect

WordNet [24] je velika leksikalna podatkovna baza v angleškem jeziku. Samostalniki, glagoli, pridevniki in prislovi so združeni v sklope kognitivnih sopomenk (angl. synsets). Vsak od sklopov izraža različen koncept. Synseti so med seboj povezani glede na konceptualno-semantične in leksikalne odnose.[24] WordNet Domains je leksikalni vir, ustvarjen na pol avtomatski način, tako da do synsetom dodali še vsaj eno domensko semantično oznako iz nabora okoli 200 oznak, ki so bile strukturirane glede na WordNet domensko hierarhijo.[14] WordNet-Affect je razširitev WordNet Domains. Vključuje podskupino synsetov, ki so primerni za predstavitev čustvenih konceptov, povezanih s čustvenimi besedami. Ti koncepti so razdeljeni v 11 čustvenih kategorij (a-label), ki jih prikazuje Tabela 2.1. [22]

A-label	**Primeri besed**
emotion	noun anger#1, verb fear#1
mood	noun animosity#1, adjective amiable#1
trait	noun aggressiveness#1, adjective competitive#1
cognitive state	noun confusion#2, adjective dazed#2
physical state	noun illness#1, adjective all in#1
hedonic signal	noun hurt#3, noun suffering#4
emotion-eliciting situation	noun awkwardness#3, adjective out of danger#1
emotional response	noun cold sweat#1, verb tremble#2
behaviour	noun offense#1, adjective inhibited#1
attitude	noun intolerance#1, noun defensive#1
sensation	noun coldness#1, verb feel#3

Tabela 2.1: Tabela čustvenih kategorij (a-label) v WordNet-Affect.

2.1.3 ANEW

ANEW [19] je kratica za Affective Norms of English Words, kar v prevodu pomeni Čustvene norme angleških besed. ANEW je nabor normativnih čustvenih ocen za veliko število besed v angleškem jeziku. Nabor je ocenjen z vidika užitka (angl. pleasure), vzburjenja (angl. arousal) in prevlade (angl. dominance) z namenom ustvariti standard za uporabo v študijah čustev in pozornosti. Na voljo je le za akademsko uporabo, za neprofitne raziskave in za nekatere učne institucije.[19]

2.1.4 AFINN

AFINN [18] je seznam angleških besed ocenjen s številom med -5 (negative) in +5 (positive). Besede so bile označene ročno. Seznam je specifično narejen za mikrobloge (Twitter) [17]. Obstajata dve različici. AFINN-111 je najnovejša verzija, ki vsebuje 2477 besed in fraz, medtem ko je AFINN-96 starejša, nepopolna verzija [18]. Primer analize z uporabo AFINN-111 lahko vidimo na Sliki 1.1 in Sliki 1.2.

2.2 Metode strojnega učenja

Metode strojnega učenja za analizo razpoloženja temeljijo na ročno označenem korpusu besed. Klasifikacija je navadno predstavljena kot dvorazredni problem z razredoma pozitiven in negativen. V nekaterih raziskavah je vključen tudi nevtralen razred, kar klasifikacijo rahlo oteži, a to ni vedno nujno. Klasifikacija subjektivnosti je v bistvu problem klasifikacije besedila. Prav zaradi tega se lahko poslužujemo obstoječih klasifikacijskih metod strojnega učenja, kot so Naivni Bayes ali metoda podpornih vektorjev (SVM). Klasifikator, naučen na ročno označenem korpusu, nato uporabimo na besedilu, ki še ni označeno. Pri tem pristopu je pomembno, kako se odločimo za lastnosti. Bolje, ko jih določimo, bolj uspešen bo klasifikator. Navadno so za vektorje z lastnostmi izbrani unigrami, posamezne besede

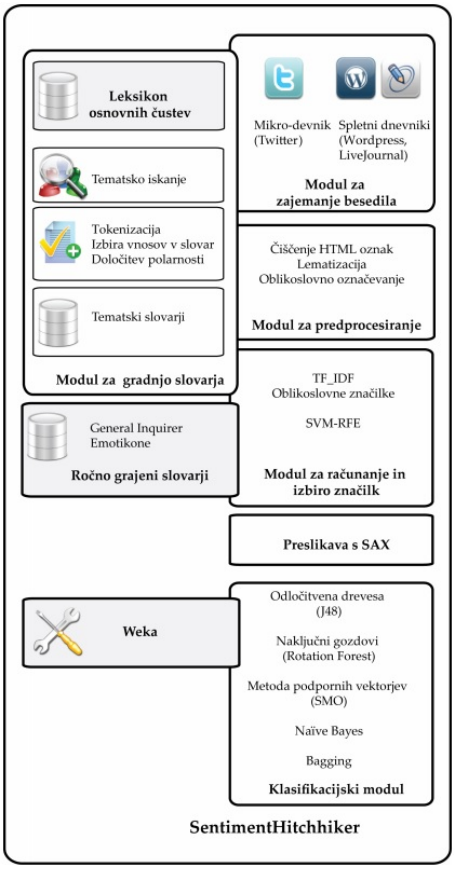
iz dokumenta, ali n-grami, dve ali več besed iz dokumenta v sekvenčnem redu, normaliziran po metodi TFIDF. Druge predlagane lastnosti vključujejo število pozitivnih besed, število negativnih besed in dolžino dokumenta. Navadno točnost teh klasifikatorjev varira med 63% in 82%, vendar so ti rezultati odvisni od lastnosti, ki smo jih izbrali.[1] Klasifikator je lahko učinkovit samo v isti domeni in na enakem tipu besedil, kot je učni korpus. Ročno označevanje besedil je zelo časovno potratno in drago.

2.3 Hibridni pristopi

Ker analiza razpoloženja sodi med kompleksnejše probleme, so se raziskovalci odločili razviti tudi hibridne sisteme, ki temeljijo na ideji odpravljanja slabosti posameznega klasifikatorja in s sinergijo več klasifikatorjev skušajo doseči boljše klasifikacijske rezultate. Eden lepih primerov, na Sliki 2.1, za angleški jezik je implementiran v orodju SentimentHitchhiker in je bil razvit v okviru doktorske disertacije dr. Mateje Verlič.[25]

2.4 Pristopi za ne-angleška besedila

Večina raziskav na področju analize razpoloženja je narejena za angleški jezik. Izgradnja sistemov za analizo razpoloženja ni lahka naloga in predvsem je zelo zamudna. Prav zato je v interesu raziskovalcev, ki se ukvarjajo z analizo razpoloženja v ne-angleških besedilih, da se poslužijo že obstoječih podatkov, korpusov, slovarjev in metod, ti pa so večinoma grajeni na angleških besedilih. Uporablja se veliko različnih pristopov. Pri leksikalnih metodah se navadno prevaja slovar ali pa besedilo, ki ga analiziramo. Prav tako lahko zgradimo leksikon za specifični jezik tako, da čim bolj natančno prevedemo le semena besed in na podlagi teh zgradimo slovar. Pri metodah strojnega učenja za analizo razpoloženja se navadno prevedejo označeni podatki, na katerih naučimo klasifikatorje ali pa besedilo, ki ga analiziramo s klasifikatorjem.[16]



Slika 2.1: Moduli orodja SentimentHitchhiker dr. Mateje Verlič.

Poglavje 3

Orodja in tehnologija

V tem poglavju opišemo orodja in tehnologije s katerimi smo si pomagali pri izdelavi diplomske naloge. Opisani so uporabljeni jeziki za izdelavo spletnih strani HTML in PHP in JavaScript, orodje za delo s podatkovnimi bazami Microsoft SQL Server in Microsoft SQL Server Management Studio in orodje za lematizacijo besedil LemmaGen. Poleg opisanega smo za izdelavo diplomskega dela uporabili tudi web server IIS, Microsoft Visual Studio, programski jezik Python in Microsoft Excel.

3.1 HTML

HTML [2] je kratica za Hyper Text Markup Language. Je označevalni jezik, ki ga uporabljamo za izdelavo dokumentov na medmrežju. HTML definira strukturo in videz spletne strani, s pomočjo oznak in atributov. HTML oznake so ključne besede, ki se nahajajo med dvema kotnima oklepajema. Pomen in funkcijo najpomembnejši oznak razložimo s pomočjo primera na Sliki 3.1.[26]

- DOCTYPE definira HTML kot tip dokumenta
- Besedilo med oznako <html> in </html> opisuje HTML dokument

- Besedilo med oznako <head> in </head> navadno podaja informacije o dokumentu
- Besedilo med oznako <title> in </title> poskrbi za ime dokumenta
- Besedilo med oznako <body> </body> opisuje vsebino, ki je prikazana na spletni strani
- Besedilo med <h1> in </h1> opisuje naslov
- Besedilo med in opisuje odstavek

Slika 3.1: Primer HTML kode.

```
<!DOCTYPE html>
<html>
<head>
  <title>Page Title</title>
</head>

<body>

  <h1>This is a Heading</h1>

  <p>This is a paragraph.</p>

</body>

</html>
```

3.2 PHP

PHP [6] je odprtokodni skriptni programski jezik, ki je posebno primeren za razvoj dinamičnih, interaktivnih spletnih strani in ga lahko vključimo tudi

v html dokumente. Zaradi dejstva, da je PHP izredno enostaven za novega uporabnika, a ponuja tudi ogromno naprednih možnosti za profesionalne programerje, je uporaba programskega jezika PHP zelo razširjena. Značilnost jezika PHP je, da se izvršitev kode vrši na strežniški strani, kjer se zgenerira HTML, ki je nato poslan odjemalcu. Odjemalec prejme rezultate skripte, kode, ki pa jih je zgenerirala, pa ne vidi. Tako lahko prikazujemo različno vsebino spletne strani glede na to kateri uporabnik je na stran prijavljen. Čeprav je razvijanje z jezikom PHP osredotočeno na skripte na strežniški strani, lahko z njim naredimo praktično karkoli. PHP ni najboljši jezik za namizne aplikacije, a lahko s poglobljenim znanjem in razširitvijo PHP-GTK ustvarjamo tudi te. PHP lahko teče na večini operacijskih sistemov in ima podporo za večino spletnih strežnikov. Prav tako podpira širok spekter podatkovnih baz [6].

Preprost primer sintakse na Sliki 3.2 prikazuje uporabo programskega jezika PHP znotraj dokumenta HTML [28]:

Slika 3.2: Primer PHP kode

```
<!DOCTYPE html>
<html>
<body>

<?php
echo "My first PHP script!";
?>

</body>
</html>
```

3.3 JavaScript

JavaScript [27] je programski jezik, ki se uporablja za interaktivnost in dinamičnost spletnih strani. Za razliko od programskega jezika PHP JavaScript teče na strani odjemalca. Navadno se odločamo, kateri jezik od teh dveh bomo izbrali tako, da se vprašamo, kdaj želimo, da se naša koda izvrši. Če se lahko izvede preden se stran naloži, lahko uporabimo PHP, če pa se mora izvršiti po tem, ko je stran naložena, uporabimo JavaScript. Podpora za JavaScript je vgrajena v vse popularnejše spletne brskalnike. Pod pogojem, da odjemalec uporablja spletni brskalnik, ki podpira JavaScript in če je nastavitev omogočena (privzeta nastavitev za to je omogočeno), potem bo spletna stran, ki uporablja JavaScript, delovala brez problemov. Tako kot PHP lahko kodo JavaScript pišemo direktno v datoteko HTML, ali pa skripto zapišemo v posebno datoteko (končnica .js) in jo nato povežemo s datoteko HTML z uporabo oznake <script> in </script>. Isto skripto lahko vključimo večkrat [27]. Primer sintakse, ki prikazuje JavaScript funkcijo vključeno v dokument HTML, je na sliki 3.3. Funkcijo shranimo med oznaki head in script in jo nato pokličemo ob kliku na gumb.

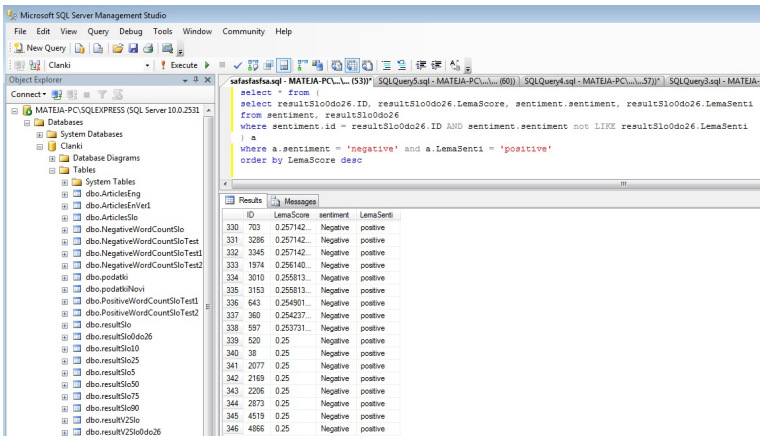
Slika 3.3: Primer JavaScript kode

```
<!DOCTYPE html>
<html>
<head>
<script>
function myFunction() {
    document.getElementById("demo").innerHTML = "Paragraph
        changed.";
}
</script>
</head>
<body>
<h1>My Web Page</h1>
```

```
<p id="demo">A Paragraph</p>
<button type="button" onclick="myFunction()">Try it</button>
</body>
</html>
```

3.4 Microsoft SQL Server

Strežnik Microsoft SQL [3] je sistem za upravljanje relacijskih zbirk podatkov. Izdelalo ga je podjetje Microsoft. Njegov osnovni programski jezik je Transact-SQL, ki je implementacija standarda ANSI/ISO SQL. Strežnik SQL lahko uporabljamo za male kot tudi velike zbirke podatkov. Do serverja dostopamo programsko ali z orodjem Microsoft SQL Server Management Studio. Na sliki 3.1 je prikazana poizvedba (angl. querry) v jeziku SQL v programu Microsoft SQL Server Management Studio.



Slika 3.1: Poizvedba v jeziku SQL.

3.5 LemmaGen

LemmaGen [9] je projekt, katerega cilj je standardizirana odprtokodna platforma za lematizacijo. Narejen je bil v okviru diplomskega dela v sodelo-

vanju z ustanovo institut "Jožef Stefan"v Ljubljani. Ta projekt se je začel zaradi pomanjkanja visoko kakovostnega orodja za lematizacijo v slovenskem jeziku. Trenutno obstaja ne le lematizator za slovenski jezik, temveč tudi 11 drugih evropskih jezikov. Sistem se je prav tako zmožen naučiti lematizacijska pravila za nove jezike iz obstoječih primerov v obliki para (besedna oblika - lema). Orodje je brezplačno, ima odprtokodno licenco za celotno kodo vključeno v projekt. Trenutno podpira 12 različnih jezikov, lematizacija pa se ne opira na strukturo stavka torej jo lahko izvajamo neodvisno, na vsaki besedi posebej. Različne API omogočajo, vključitev LemmaGen v različne projekte. Viri so na voljo v več izvedbah (C++, C++.NET, Python in C#.NET). Podpirajo več različnih platform. Orodje je zelo dobro dokumentirano, implementacija pa je zelo učinkovita in hitra (hitrost procesiranja je 1 miljon besed na sekundo). Pohvalijo se lahko tudi s kakovostjo vnaprej zgrajenih lematizacijskih modelov. [8] V diplomskem delu uporabimo verzijo orodja LemmaGen - LemmaSharp za programski jezik C#. Na sliki 3.2 pa je prikazano delovanje LemmaSharp - lematizacija besed v enem odstavku.

Slika 3.2: Primer LemmaSharp - lematizacija besed v enem odstavku.[8]

Poglavje 4

Ročno ocenjevanje člankov

Besedilni korpusi [29] so obsežne zbirke besedil v naravnem jeziku, zajete v določenem obdobju iz množičnih medijev, shranjene v strukturirani obliki in s pomočjo jezikovnih tehnologij opremljene z označbami. V tem poglavju opišemo, kako smo izdelali korpus slovenskih novic, opremljen z označbami, ki nam povedo subjektivnost posameznega besedila.

Izdelali smo enostavno spletno aplikacijo za ročno ocenjevanje člankov. Pri izdelavi je bila uporabljena tehnologija Microsoft SQL, programski jezik PHP in JavaScript. Zbrali smo slovenske novice s spleta iz obdobja dveh tednov, od 15.3.2013 do 31.3.2013. Članki so bili zajeti s spletnih strani, kot so 24ur.si, Žurnal24.si, delo.si, slovenskenovice.si in podobne.

Članki so bili podani v obliki datotek xml. Vsaka datoteka je vsebovala več člankov. Vsak članek je poleg besedila vseboval veliko drugih informacij, kot na primer na kateri spletni strani se članek nahaja in pa svoj identifikacijski niz, po katerem smo lahko ločevali med članki. Podatke iz datotek xml smo s pomočjo poizvedb SQL prenesli v vnaprej pripravljeno bazo podatkov. Narejeni so bili tudi uporabniški profili, s katerimi so se ocenjevalci vpisali na spletno stran. Skozi profile smo lahko nadzorovali količino ocenjenih člankov glede na posameznika.

Aplikacija deluje tako, da najprej uporabnik vpiše uporabniško ime in geslo, nato se aplikacija poveže s podatkovno bazo in v spletnem brskalniku

prikaže spletno stran, na kateri se nahaja prvi neocenjen članek. Pod oknom s spletno stranjo se prikaže besedilo članka, kajti nekatere strani so bile od prve objave premaknjene oziroma postale nedosegljive. Uporabnik lahko članek oceni s tremi med seboj izključujočimi se možnostmi. Članek je lahko, glede na ocenjevalčevo osebno mnenje, ocenjen kot pozitiven, negativen ali nevtralen. Šele, ko je izbrana ena izmed podanih možnosti, se ocenjevalec lahko pomakne na naslednji neocenjeni članek v bazi.

Pri izdelavi spletne aplikacije smo pri testiranju naleteli na nekaj težav. Prva izmed težav je bilo ocenjevanje člankov več uporabnikov hkrati. Ker aplikacija deluje tako, da po vpisu v aplikacijo pokaže prvi neocenjeni članek v bazi, bi lahko nastala situacija, kjer se je uporabnik A vpisal, dobil za ocenjevanje članek X in preden ga je ocenil, se je lahko vpisal tudi uporabnik B in dobil v oceno isti članek X. Če bi v tej situaciji uporabnik A ocenil članek, in nato uporabnik B ocenil isti članek, bi bila ocena uporabnika A prepisana z oceno uporabnika B. Ker smo najprej načrtovali pet ocenjevalcev, bi bil lahko to velik problem. A se je na koncu izkazalo, da smo imeli le dva aktivna ocenjevalca, zato smo problem rešili tako, da smo ocenjevali izmenično. Če bi imeli več aktivnih ocenjevalcev, bi morali problem rešiti z dodatno tabelo, ki bi hranila podatke o ocenjevalcu, članku in oceni. Ker aplikacija ni bila glavni izdelek tega dela in smo le želeli hitro pridobitev ocenjenih člankov, smo to težavo lažje rešili z izmeničnim ocenjevanjem.

Druga napaka, na katero smo naleteli, je bila povezana s kompatibilnostjo aplikacije z različnimi spletnimi brskalniki. Aplikacijo smo testirali v dveh najbolj uporabljenih brskalnikih, Google Chrome in Mozilla Firefox. Težava se je pojavila le v brskalniku Mozilla Firefox, saj je bila namesto spletne strani članka prikazana prazna stran s sporočilom o napaki. Ker do te napake ni prišlo v spletnem brskalniku Google Chrome, smo se odločili, da ocenjevanje izvajamo le v tem brskalniku. Na druge napake v aplikaciji nismo naleteli. Skupno je bilo ročno ocenjenih 5000 spletnih člankov slovenskih novic. Slika 4.1 prikazuje aplikacijo za ročno ocenjevanje člankov.

Slika 4.1: Delovanje aplikacije za ročno ocenjevanje člankov.

Poglavje 5

Izdelava slovenskega slovarja razpoloženja

V tem poglavju opišemo izdelavo slovenskega slovarja razpoloženja in ovrednotimo njegovo kakovost. Najprej opišemo uporabo slovarja v klasifikatorju razpoloženja članka in mere, ki smo jih uporabili za ocenjevanje kakovosti slovarja. Nato opišemo postopek izdelave slovarja, ugotovimo njegove pomanjkljivosti in uvedemo izboljšave. Vse primerjamo z večinskim klasifikatorjem in z alternativno metodo, ki prevede besedila v angleščino in za klasifikacijo razpoloženja uporabi originalni (angleški) slovar.

5.1 Izdelava prve verzije slovarja - Alfa

Osnovni seznam besed, iz katerega izhajamo, je osnovan na angleških seznamih besed General Inquirer [7] dveh največjih kategorij, Positiv in Negativ.

Oznaka	**Št. besed**	**Primer besed**
Negativen	2293	arrest, bankruptcy, fool, explosion, drugs
Pozitiven	1914	goal, smart, gift, genious, integrity, honesty

Tabela 5.1: Primeri angleških besed iz seznamov General Inquirer Negativ in Positiv, ki izražajo subjektivnost.

Lastnosti seznama so prikazane v Tabeli 5.1. Paket vsebuje dve tekstovni datoteki. TAGneg.txt vsebuje seznam 2293 negativnih besed in TAGpos.txt seznam 1914 pozitivnih besed.

Za prevajanje smo uporabili tri prosto dostopna orodja. Google Translate[1], slovarje Pons[2], in angleško angleški slovar The Free Dictonary[3]. Avtomatsko prevedene besede smo ročno preverili in popravili napake.

Najprej smo zaradi želje po enostavnosti kopirali oba seznama besed v Google Translate in dobili seznama besed, ki naj bi bila v slovenskem jeziku. Po temeljitem pregledu prevedenih seznamov smo ugotovili, da temu ni tako. V TAGPos od 1914 besed 541 besed ni bilo prevedenih in so ostale v angleškem jeziku. Le 205 besed je bilo pravilno prevedenih in so se pojavile v končnem seznamu prevedenih besed. V napačnem jeziku je bilo 19 besed. Preostale besede so bile v napačni obliki ali narobe prevedene.

V TAGNeg od 2293 besed 638 ni bilo prevedenih. 264 besed je bilo pravilno prevedenih in so se pojavile na končnem seznamu. Besed v tujem jeziku je bilo 46. Nepravilne besede smo ročno popravili s pomočjo slovarja Pons. Dodali smo tudi sopomenke, če so obstajale. Dodajanje sopomenk smo omejili na maksimalno 3 besede. V pozitivnem seznamu besed smo dodali sopomenke 553 besedam, negativnemu pa 246 besedam.

Prevedena seznama besed vsebujeta 2555 pozitivnih besed in 2562 negativnih besed. Seznama smo dodatno uredili s pomočjo programskega jezika Python. Odstranili smo dvojnike, vsako besedo postavili v novo vrstico, poiskali fraze in seznama uredili po abecedi. Na seznamu pozitivnih besed je bilo 717 odstranjenih ponovljenih besed in najdenih 36 bigramov. Na seznamu negativnih besed je bila odstranjena 601 ponovljena beseda, najdenih pa je bilo 18 bigramov. Šumnikov nismo odstranjevali, ker jih lažje kasneje odstranimo, kot pa dodamo, če so potrebni. Python besede uredi tako, da postavi vse šumnike na konec. Zato smo sezname ročno popravili, da je bil zagotovljen abecedni vrstni red. Slovnico in črkovanje smo preverili s pro-

[1]Google Translate: https://translate.google.com/

[2]Slovarji Pons: http://www.pons.si/

[3]The Free dictionary: http://www.thefreedictionary.com/

gramom Microsoft Office Word. Končni seznam pozitivnih besed, primer v Tabeli 5.2, vsebuje 1838 besed, seznam negativnih pa 1961 besed.

Google Translate se je izkazal kot slabo orodje za prevajanje velikega števila posameznih besed. Ker je prosto dostopen in lahko kdorkoli ureja, dodaja ter predlaga prevode, velikokrat najdemo povsem neprimerne prevode. Primeren je za hitro in površno prevajanje v kontekstu, vendar ne za veliko količino posameznih besed. Slovarji Pons so se bolje izkazali, saj so poleg prevodov ponudili tudi razlage in obliko besede. Vendar je tu treba izpostaviti dejstvo, da smo z njimi prevajali besedo za besedo, kar je bilo veliko časovno potratnejše, medtem ko smo z orodjem Google Translate prevedli celoten seznam besed. Kljub dobremu delovanju slovarjev Pons je v njih manjkalo 37 besed iz obeh seznamov. Navadno so bile to posebne besede v širši rabi kot *'amour'* ali izpeljanke kot *'beauteous'* in pa staroangleške besede. Nekaj je bilo tudi izrazov povezanih z medicinskimi pripomočki. Pomen teh smo poiskali s pomočjo prosto dostopnega angleško angleškega slovarja The Free Dictionary in pripisali ustrezno slovensko besedo glede na angleško razlago. Če beseda ni bila najdena tam, smo jo izpustili.

Oznaka	**Št. besed**	**Primer besed**
Negativen	1961	aretacija, bankrot, bedak, eksplozija, droga
Pozitiven	1838	cilj, pameten, dar, genij, integriteta, iskrenost

Tabela 5.2: Primeri pozitivnih in negativnih besed iz slovarja Alfa za slovenski jezik.

5.2 Ocenjevanje kakovosti slovarja

Kakovost slovarja ocenjujemo posredno preko klasifikatorja, ki ga zgradimo po zgledu leksikalne analize razpoloženja iz članka More than words: Quantifying Language to Measure Firms' Fundamentals (Več kot besede: Merjenje jezika za merjenje temeljev podjetja) [23]. Klasifikator klasificira v tri razrede: pozitiven, nevtralen in negativen, zato so vse mere za oceno

kakovosti klasifikatorja izpeljane iz matrike zmot velikosti 3x3. Ker je napaka iz pozitivnega v negativen razred (in obratno) hujša kot napaka iz pozitivnega ali negativnega v nevtralen razred, za ocenjevanje kakovosti klasifikatorja uporabimo tudi mere ordinalne regresije opisane v člankih [5] in [10].

5.2.1 Opisi mer

Za ocenjevanje uspešnosti klasifikacije s slovarjem uporabimo naslednje mere in metode:

Matrika zmot

Matrika zmot [12] (confusion matrix) je matrika števila ali odstotkov napačnih klasifikacij.

Tabela 5.3 predstavlja matriko zmot za trirazredni klasifikacijski problem. Na diagonali matrike so podana števila pravilnih klasifikacij. Vsota vsake vrstice podaja število oziroma delež primerov pravih razredov, vsote stolpcev pa nam povejo število ali delež primerov, ki so klasificirani v posamezen razred. Iz odstopanja dveh vsot sklepamo na pristranskost klasifikatorja [12]. Iz Tabele 5.3 lahko razberemo pomen posameznega polja matrike za trirazredni problem z razredi A, B in C.

	napovedani razred			
pravi razred	razred A	razred B	razred C	vsota
razred A	AA	AB	AC	VA
razred B	BA	BB	BC	VB
razred C	CA	CB	CC	VC
vsota	KA	KB	KC	N

Tabela 5.3: Matrika zmot za trirazredni problem.

V matriki zmot, polje XY pomeni število primerov razreda X, ki jih je klasifikator uvrstil v razred Y. Na primer AB je število primerov razreda A, ki jih je klasifikator uvrstil v razred B. Polje VX nam pove vsoto vrstice

razreda X ali število primerov razreda X. Na primer VA = AA + AB + AC je vsota vrstice razreda A ali število primerov razreda A. Polje KX nam pove vsoto stoplca razreda X ali število primerov, ki jih je klasifikator uvrstil v razred X. Na primer KA = AA + BA + CA je vsota stolpca razreda A ali število primerov, ki jih je klasifikator uvrstil v razred A. Polje N = AA + AB + AC + BA + BB + BC + CA + CB + CC nam pove število vseh primerov.

Klasifikacijska točnost

Klasifikacijska točnost [12] (classification accuracy) nam pove delež pravilno klasificiranih primerov. Definirana je z enačbo:

$$T = \frac{P}{N} \times 100\% = \frac{AA + BB + CC}{N} \times 100\%$$

kjer je P število pravilno klasificiranih primerov (v Tabeli 5.3 je P = AA + BB + CC) in N število vseh primerov. V primeru klasifikacijske točnosti je vsako nestrinjanje med klasifikatorjem in anotatorjem uteženo z 1, kar prikažemo z matriko tež napak v Tabeli 5.4 Za pravilne rezultate se štejejo le klasifikacije v pravilne razrede, vse napačne klasifikacije štejejo za enakovredno napako.

	napovedani razred		
pravi razred	razred A	razred B	razred C
razred A	0	1	1
razred B	1	0	1
razred C	1	1	0

Tabela 5.4: Matrika teže napak za klasifikacijsko točnost. Na diagonali se nahajajo pravilne klasifikacije, zato je utež napake enaka 0. Vse ostale klasifikacije štejejo za enako hude (težke) napake.

Mere ordinalne regresije

Srednja absolutna napaka (mean absolute error, MAE) in **srednja kvadratna napaka** (mean squared error, MSE) sta meri ordinalne regresije,

ki bolj kaznujeta napake v bolj oddaljene razrede. Meri predpostavljata urejenost razredov in enako razdaljo med razredi. V našem primeru je razdalja med sosednjima razredoma 1, razredi pa so urejeni v vrstnem redu A > B > C. Napaka MAE, katere matrika teže napak je prikazana v Tabeli 5.5, kaznuje večje napake (med razredoma A in C) dvojno. Napaka MSE, katere matrika teže napak je prikazana v Tabeli 5.6, kaznuje večje napake kvadratično z razdaljo med razredoma, kar je v našem primeru štirikratno. Torej je MSE boljša v situacijah kjer imamo manjšo toleranco za večje napake. [10].

	napovedani razred		
pravi razred	razred A	razred B	razred C
razred A	0	1	2
razred B	1	0	1
razred C	2	1	0

Tabela 5.5: Matrika teže napak za srednjo absolutno napako. Na diagonali se nahajajo pravilne klasifikacije, zato je utež napake enaka 0. Napake iz razreda A v B so manjše kot napake iz razreda A v C in obratno.

	napovedani razred		
pravi razred	razred A	razred B	razred C
razred A	0	1	4
razred B	1	0	1
razred C	4	1	0

Tabela 5.6: Matrika teže napak za srednjo kvadratno napako. Na diagonali se nahajajo pravilne klasifikacije, zato je utež napake enaka 0. Vse vrednosti so kvadrirane teže absolutnih napak iz Tabele 5.5.

MAE in MSE izračunamo kot Hadamardov produkt matrike zmot in matrike teže napak, vrednosti v dobljeni matriki seštejemo in delimo s številom primerov. Hadamardov produkt dveh matrik je matrika, ki jo dobimo tako da zmnožimo istoležne elemente. Za dve matriki A in B istih dimenzij m×n

je Hadamardov produkt

$$(A \circ B)_{i,j} = (A)_{i,j} \cdot (B)_{i,j} \text{ [31].}$$

Na primer srednjo absolutno napako tako izračunamo v našem primeru po enačbi $MAE = \frac{AB+BC+BA+CB+2AC+2CA}{N}$.

Točnost znotraj n, kjer je n = 1, T1 [5] (accuracy within n, where n = 1, ACC1). Točnost znotraj n predstavlja družino mer, podobno točnosti, le da dovoljuje večjemu razponu rezultatov, da se uvrstijo kot pravilni. V primeru kjer, imamo ocene od 1 do 5, lahko s točnostjo znotraj 1, pri pravilnem razredu 5, upoštevamo kot pravilne napovedi v razred 5 in razred 4. Točnost znotraj 0 je klasifikacijska točnost. V diplomskem delu si s točnostjo znotraj 1 pomagamo pri merjenju večjih napak. Za napako štejemo le, če klasifikator klasificira pozitivne primere kot negativne in obratno. Če klasifikator klasificira pozitiven primer kot nevtralen ali negativen primer kot nevtralen, to ni napaka. V celoti štejemo kot hujšo napako klasifikacijo negativnega članka kot pozitivnega in obratno. Večji T1 pomeni boljši klasifikator. Napake v primeru točnosti znotraj 1, v matriki zmot utežimo, kot je to prikazano v Tabeli 5.7 in jo izračunamo, tako kot MAE in MSE, s pomočjo Hadamardovega produkta, v našem primeru kot [5] $T1 = 1 - \frac{AC+CA}{N}$.

	napovedani razred		
pravi razred	razred A	razred B	razred C
razred A	0	0	1
razred B	0	0	0
razred C	1	0	0

Tabela 5.7: Za napako se štejejo le napačne klasifikacije razreda A v razred C in klasifikacije razreda C v razred A.

Priklic in preciznost

Preciznost (precision) [12] je mera točnosti, ki nam pove odstotek pravilno klasificiranih primerov določenega razreda, glede na število klasificiranih

primerov v ta določen razred. Glede na Tabelo 5.3 jo za razred A izračunamo kot

$$PreciznostA = \frac{AA}{AA + BA + CA} = \frac{AA}{KA}.$$

Priklic (recall) [12] je mera, ki nam pove, kako sposoben je klasifikator izbrati primere določenega razreda izmed množice primerov tega določenega razreda. Glede na Tabelo 5.3 ga za razred A izračunamo kot

$$PriklicA = \frac{AA}{AA + AB + AC} = \frac{AA}{VA}.$$

Mera F1

Mera F1 [30] (F1-measure) ali uravnotežena mera F (balanced F-measure) je harmonična sredina priklica in preciznosti. Harmonična sredina (harmonic mean) je v matematiki ena od srednjih vrednosti. Harmonična sredina H [32] dveh pozitivnih realnih števil a in b je določena kot:

$$H(a, b) = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2ab}{a + b}.$$

Mero F1 izračunamo po enačbi

$$F1 = 2 \times \frac{Priklic \times Preciznost}{Priklic + Preciznost}.$$

Mera $\mathbf{F1_{avg}}$ [10], katere enačba je

$$F1_{avg} = \frac{F1\text{pozitiven} + F1\text{negativen}}{2},$$

je povprečje mere F1 za pozitiven in negativen razred. To mero uporabljajo kot mero za kvaliteto kakovosti klasifikatorjev za analizo razpoloženja na tekmovanjjih analize razpoloženja SemEval [21]. Večja kot je mera F1 oziroma $F1_{\text{avg}}$, boljši je klasifikator.

Večinski klasifikator

Točnost večinskega klasifikatorja je mera, ki nam pove, kakšna je točnost v primeru, da klasifikator uvrsti vse primere v večinski razred. V našem

primeru je to razred negativen. V tabeli 5.8 je podana matrika zmot za večinski klasifikator na našem primeru. V Tabeli 5.9 so podane izračunane vrednosti mer za ocenjevanja klasifikatorja.

	napovedani razred			
pravi razred	pozitiven	nevtralen	negativen	vsota
pozitiven	0	0	1349	1349
nevtralen	0	0	1603	1603
negativen	0	0	2048	2048
vsota	0	0	5000	5000

Tabela 5.8: Matrika zmot za večinski klasifikator. Ker je večinski razred negativen, večinski klasifikator klasificira vse primere negativno.

mere	vrednost
točnost	0,41
MAE	0,86
MSE	1,4
mera $F1_{avg} = \frac{F1 pozitiven + F1 negativen}{2}$	0,29
T1 (ACC1)	0,73

Tabela 5.9: Rezultati mer za ocenjevanje večinskega klasifikatorja.

5.2.2 Opis klasifikatorja s slovarjem

Za leksikalno analizo izdelamo klasifikator, ki avtomatsko oceni subjektivno usmerjenost besedila. Postopek avtomatskega ocenjevanja je bil za vse članke enak. Pri izdelavi klasifikatorja smo uporabili orodje LemmaGen [8] Multilingual Open Source Lemmatisation Framework, programski jezik C# in program Visual Studio. Program uporabi LemmaGen in z njim lematizira vhodno besedilo in slovar, ter prešteje posamezne pozitivne in negativne besede, ki se nahajajo v besedilu. Nato izračuna vrednost subjektivne us-

merjenosti za celotno besedilo po formuli:

$$s = \frac{N\text{pozitivnih\_besed} - N\text{negativnih\_besed}}{N\text{pozitivnih\_besed} + N\text{negativnih\_besed}}$$

pri čemer je N število pojavitev določene besede iz slovarja v besedilu. Rezultat je vrednost med 1 in -1. Glede na izbrani mejni vrednosti za pozitivnost in negativnost nato določimo, ali je vrednost subjektivne usmerjenosti pozitivna, negativna ali nevtralna. Pridobljene podatke smo zapisovali v bazo podatkov na lokalnem strežniku. Osnovni vzorec klasifikatorja leksikalnega pristopa, ki je uporabljen v tem delu, lahko povzamemo v petih korakih:

1. Predprocesiranje besedila na katerem vršimo analizo (lematizacija, odstranjevanje ločil in tekstovnih oznak).

2. Nastavitev števca pozitivnih besed $p = 0$ in števca negativnih besed $n = 0$.

3. Za vsako lemo preveri, če se nahaja v slovarju subjektivnih besed.

   Če se lema nahaja v slovarju,

   i. Če je lema pozitivna, $p = p + 1$

   ii. Če je lema negativna, $n = n + 1$

4. Izračunaj subjektivno usmerjenost s za celotno besedilo s po enačbi

   $$s = \frac{p - n}{p + n} \text{ , če } p + n > 0, \text{ sicer } s = 0.$$

5. Poglej skupno subjektivno usmerjenost za celotno besedilo s in določi mejne vrednosti za pozitivno m_p in negativno m_n:

   (a) Če $s \geq m_p$, besedilo je pozitivno.

   (b) Če $s \leq m_n$, besedilo je negativno.

   (c) Če $m_n < s < m_p$, besedilo je nevtralno.

5.3 Prevedeni članki in originalni slovar

Analizo razpoloženja za slovenski, kot tudi za ostale ne-angleške jezike, lahko naredimo tako, da besedilo prevedemo v angleščino in uporabimo metode, ki so razvite za angleški jezik [16]. Slovenske novice smo s pomočjo orodja Google Translate prevedli v angleščino. Kot slovar smo uporabili angleški slovar General Inquirer [7], seznama Positiv in Negativ in klasifikacijo, opisano v prejšnjem razdelku. Rezultati mer za ocenjevanje so v Tabeli 5.12.

mejne vrednosti	točnost
$m_p = m_n = 0$	0,43
$m_p = 0,26,\ m_n = 0$	0,43
$m_p = 0,10,\ m_n = 0,10$	0,42
$m_p = 0,25,\ m_n = 0,25$	0,39
$m_p = 0,50,\ m_n = 0,50$	0,35
$m_p = 0,75,\ m_n = 0,75$	0,33

Tabela 5.10: Klasifikacijske točnosti za različne mejne vrednosti klasifikatorja, ki uporablja originalni angleški slovar in mu podamo v angleščino prevedene članke.

	napovedani razred			
pravi razred	pozitiven	nevtralen	negativen	vsota
pozitiven	631	468	250	1349
nevtralen	568	587	448	1603
negativen	439	638	971	2048
vsota	1638	1693	1669	5000

Tabela 5.11: Matrika zmot za klasifikator, ki uporablja originalni angleški slovar, z mejnimi vrednostmi $m_p = 0,26$, $m_n = 0$.

mere	vrednost
točnost	0,43
MAE	0,70
MSE	0,98
mera $F1_{avg} = \frac{F1pozitiven+F1negativen}{2}$	0,48
T1 (ACC1)	0,86

Tabela 5.12: Rezultati mer za ocenjevanje klasifikatorja, ki uporablja originalni angleški slovar.

Najprej smo poskušali dobiti čim večjo točnost samo s premikanjem mejnih vrednosti. Tabela 5.10 prikazuje klasifikacijsko točnost klasifikatorja za različne mejne vrednosti za pozitvno m_p in negativno m_n. Najboljša vrednost je 0,43 pri vrednostih parametrov $m_p = 0,26$ in $m_n = 0$. z večanjem parametrov m_p in m_n se točnost manjša. Če primerjamo izračunane vrednosti za klasifikator s prevedenimi članki in originalnim angleškim slovarjem ter večinski klasifikator opazimo, da naš klasifikator deluje malce bolje kot večinski. Točnost je 0,02 večja kot pri večinskem klasifikatorju. Napaki MAE in MSE se znižata za 0,16 in 0,42. Mera $F1_{avg}$ se zviša za 0,19 in mera T1 za 0,13. Rezultati so v primerjavi z večinskim klasifikatorjem boljši, a ne za veliko.

5.4 Preveden slovar Alfa

Klasificirali smo članke v slovenskem jeziku z uporabo prevedena seznama pozitivnih in negativnih besed, slovarja Alfa, opisanega v razdelku 5.1 in klasifikatorjem iz razdelka 5.2.2. V Tabeli 5.13 se nahajajo rezultati glede na izbrani mejni vrednosti. Najboljša vrednost je 0,35 pri vrednostih parametrov $m_p = 0,26$ in $m_n = 0$. Točnost je za 0,06 nižja kot pri večinskem klasifikatorju. Z večanjem parametrov m_p in m_n se tudi v tem primeru točnost manjša.

mejne vrednosti	točnost
$m_p = m_n = 0$	0,34
$m_p = 0,26,\ m_n = 0$	0,35
$m_p = 0,10,\ m_n = 0,10$	0,33
$m_p = 0,25,\ m_n = 0,25$	0,32
$m_p = 0,50,\ m_n = 0,50$	0,32
$m_p = 0,75,\ m_n = 0,75$	0,31

Tabela 5.13: Klasifikacijske točnosti za različne vrednosti parametrov m_p in m_n za klasikator s slovarjem Alfa.

Tabela 5.14 prikazuje matriko zmot za mejni vrednosti $m_p = 0,26$ in $m_n = 0$, kjer klasifikator dosega največjo klasifikacijsko točnost 0,35. V Tabeli 5.15 so prikazane izračunane vrednosti mer za ocenjevanje.

	napovedani razred			
pravi razred	pozitiven	nevtralen	negativen	vsota
pozitiven	870	387	92	1349
nevtralen	949	464	190	1603
negativen	948	707	393	2048
vsota	2767	1558	675	5000

Tabela 5.14: Matrika zmot za klasifikator s slovarjem Alfa, z mejnimi vrednostmi $m_p = 0,26,\ m_n = 0$.

mere	vrednost
točnost	0,35
MAE	0,86
MSE	1,28
mera $\mathrm{F1}_{\mathrm{avg}} = \frac{F1\mathrm{pozitiven}+F1\mathrm{negativen}}{2}$	0,29
T1 (ACC1)	0,79

Tabela 5.15: Rezultati mer za ocenjevanje klasifikatorja s slovarjem Alfa.

Če pogledamo Tabelo 5.14, matriko zmot za klasifikator s slovarjem Alfa, se nam iz rezultatov zdi, da klasificiranje v razred pozitiven deluje relativno dobro, medtem ko za ostala dva razreda ne dobimo dobrih rezultatov. Razlogi za to so najbrž v tem, da je seznam pozitivnih besed v slovenskem jeziku bolje preveden in vsebuje več bolj pomembnih in močnih subjektivnih besed na področju novic v primerjavi s seznamom negativnih besed. To lahko opazimo, če pogledamo Tabelo 5.11, saj je tam delež pravilnih negativno klasificiranih člankov v primerjavi z napačnimi višji kot v Tabeli 5.14.

Z začetnimi rezultati klasifikatorja s prevednim slovarjem Alfa in originalnimi članki nismo bili zadovoljni. Sploh ne s kvalifikacijo negativnega razreda, kjer je naš klasifikator pravilno odkril le 393 člankov z negativno polarnostjo. Tudi, ko smo primerjali mere za ocenjevanje iz Tabele 5.15 z merami večinskega klasifikatorja in klasifikatorja z originalnim angleškim slovarjem so bili rezultati zelo slabši. Klasifikator s slovarjem Alfa je deloval še slabše kot večinski klasifikator. Zato smo se še enkrat odločili pregledati naš slovar. Ugotoviti smo želeli, katere so besede, ki povzročajo največ napačnih klasifikacij.

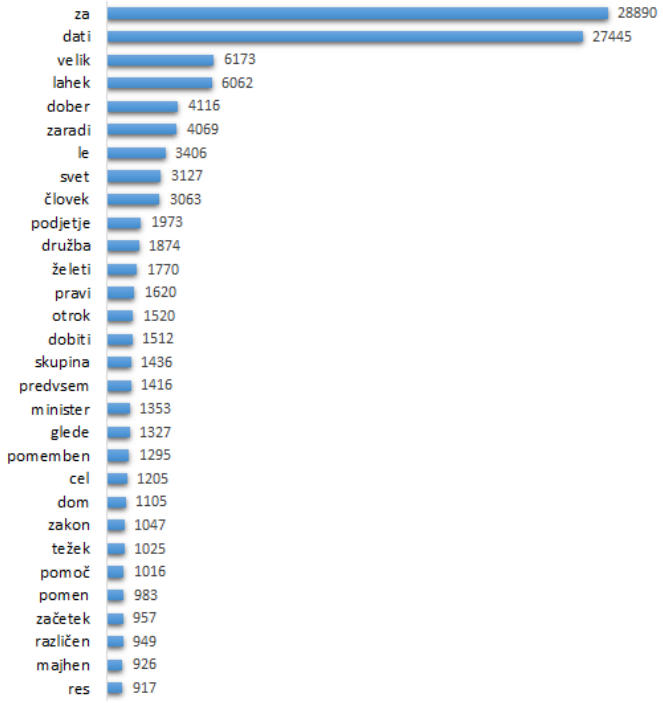
Za vsako besedo iz slovarja smo prešteli, kolikokrat se pojavi v besedilih. Zraven smo pridobili in shranili tudi število člankov, v katerih se pojavi, število člankov, ki imajo polarnost drugačno od besede, število člankov, ki se ujemajo s polarnostjo besede, in število nevtralnih člankov v katerih se ta beseda pojavi. Naredili smo dve tabeli, eno za negativne besede in eno za pozitivne besede. Iskali smo večje anomalije. Najprej smo tabeli uredili po številu besed, nato po številu člankov z nasprotno polarnostjo. Slika 5.1 prikazuje padajoče urejen seznam 30 najpogostejših pozitivnih besed. Število pojavitev besed nam pove, kako zelo beseda vpliva na izračun polarnosti besedil. Najprej smo pregledali tabelo pozitivnih besed. Takoj smo opazili prvi dve števili. Besedici *'za'* in *'dati'* sta v primerjavi z ostalimi besedami imeli ogromno število pojavitev (28890 in 27445 proti 6173, ki jima sledi na tretjem mestu). Še več, ko smo pogledali, v katerih člankih se največkrat pojavljata, smo ugotovili, da v negativnih.

```
SELECT *
  FROM [Clanki].[dbo].[PositiveWordCountSloTest2]
  order BY TotalWordCount_Slo DESC
```

Results | Messages

	ID_Pos_Slo	Word_Slo	TotalWordCount_Slo	TotalArticleCount_Slo	OppositePol_Slo	SamePol_Slo	NeutralPol_Slo
1	1681	za	28890	4018	1681	1095	1242
2	122	dati	27445	3642	1609	966	1067
3	1598	velik	6173	2414	977	685	752
4	443	lahek	6062	2357	999	645	713
5	139	dober	4116	1821	660	590	571
6	1739	zaradi	4069	1942	1026	410	506
7	449	le	3406	1755	784	454	517
8	1424	svet	3127	1504	611	420	473
9	100	človek	3063	1299	610	374	315
10	857	podjetje	1973	680	346	152	182
11	212	družba	1874	737	374	174	189
12	1829	želeti	1770	1099	446	334	319
13	988	pravi	1620	983	406	306	271
14	808	otrok	1520	478	215	157	106
15	142	dobiti	1512	1006	435	287	284
16	1299	skupina	1436	660	272	179	209
17	1016	predvsem	1416	962	403	273	286
18	507	minister	1353	451	274	85	92
19	263	glede	1327	874	424	210	240
20	895	pomemben	1295	800	284	246	270
21	72	cel	1205	836	393	231	212
22	169	dom	1105	668	289	200	179
23	1719	zakon	1047	454	280	84	90
24	1467	težek	1025	726	324	190	212
25	903	pomoč	1016	604	285	171	148
26	897	pomen	983	695	283	198	214
27	1686	začetek	957	711	300	181	230
28	1185	različen	949	621	214	193	214
29	488	majhen	926	626	256	181	189
30	1235	res	917	599	304	145	150

Slika 5.1: Najpogostejših 30 pozitivnih besed slovarja Alfa.



Slika 5.2: Histogram najpogostejših 30 pozitivnih besed slovarja Alfa, ki se pojavijo v negativnih člankih.

Nato smo seznam uredili tako, da smo pogledali katere besede imajo število negativnih člankov večje kot število pozitivnih člankov. Ugotovili smo, da je takšnih 923 besed od 1838, kar je kar polovica. Za vsako besedo na tem seznamu, smo izsledili iz katere angleške besede je bila prevedena, preverili, ali je prevod ustrezen in ali ima prevod v slovenski jezik res pozitiven pomen. Nato smo besedo odstranili, zamenjali s primernejšim prevodom ali pa jo pustili na seznamu, če smo ugotovili, da je beseda pozitivna, prevod pa ustrezen. Če smo bili v dvomih, ali je beseda res pozitivna, smo se vprašali ali je negativna, če je bil odgovor ne, smo jo pustili na seznamu, če je bil odgovor da, smo jo odstranili. Slika 5.2 prikazuje histogram besed za pozitivne besede slovarja Alfa, ki se pojavijo večkrat v negativnih člankih kot pozitivnih.

Nato smo isti postopek uporabili še na seznamu negativnih besed. Seznam negativnih besed nas je pozitivno presenetil. Ugotavljanje polarnosti besed je bilo v primeru negativnih besed veliko lažje. Spornih besed, kjer nismo vedeli ali besedo obdržati ali izpustiti, je bilo veliko manj. Slika 5.4 prikazuje histogram besed za negativne besede slovarja Alfa, ki se večkrat pojavijo v pozitivnih člankih kot negativnih. Na Sliki 5.3 že na prvi pogled opazimo, da vrtoglavih številk, kot pri pozitivnem seznamu ni. Ko smo natančneje pogledali še število člankov z nasprotno polarnostjo, smo ugotovili, da je ta le v redkih primerih večja od števila člankov z negativno polarnostjo, torej polarnostjo besede. To se je zgodilo le v 227 primerih. Najbolj očiten primer je bila beseda *'volja'*, s pojavitvijo v 124 člankih enake polarnosti in 130 člankih nasprotne polarnosti. Pregled negativnega seznama smo kljub temu izvedli enako kot pregled pozitivnega. Pregledali smo vse besede, ki so imele visoko število nasprotno označenih člankov in vse besede, ki so imele število člankov nasprotne polarnosti večje od števila člankov enake polarnosti. Zopet smo izsledili, iz katere angleške besede izhaja beseda, preverili, ali je prevod ustrezen, in ali ima slovenski prevod besede resnično negativen pomen. Nato smo besedo odstranili, zamenjali s primernejšim prevodom ali pa jo pustili na seznamu, če smo ugotovili, da je beseda negativna, prevod pa ustrezen. Rezultat revizije je slovar razpoloženja za slovenski jezik Beta.

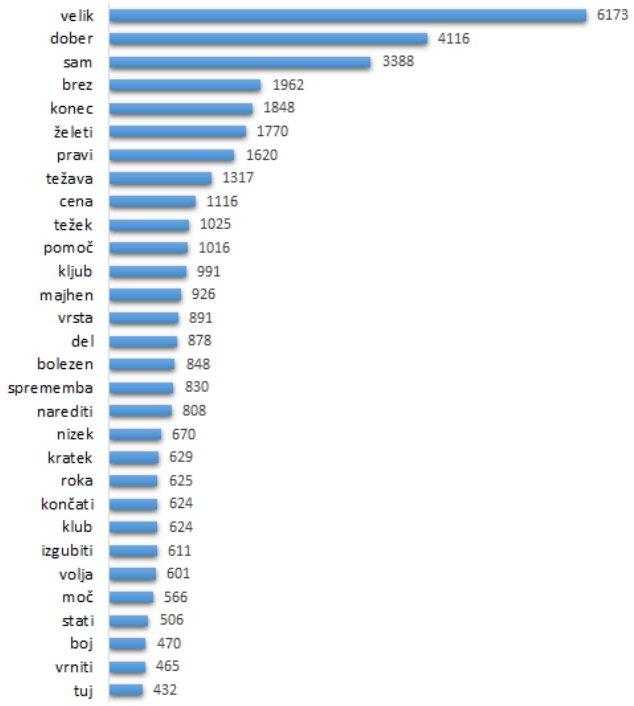
SQLQuery16.sql - MATEJA-PC\.....55))* | SQLQuery15.sql - MATEJA-PC\.....(53)) | SQLQuery14.sql - MATEJA-PC\.....52))*

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT *
  FROM [Clanki].[dbo].[NegativeWordCountSloTest2]
  order BY OppositePol_Slo DESC
```

Results | Messages

	ID_Neg_Slo	Word_Slo	TotalWordCount_Slo	TotalArticleCount_Slo	OppositePol_Slo	SamePol_Slo	NeutralPol_Slo
1	1722	velik	6173	2414	685	977	752
2	165	dober	4116	1821	590	660	571
3	1399	sam	3388	1649	421	784	444
4	1952	želeti	1770	1099	334	446	319
5	416	konec	1848	1175	310	530	335
6	87	brez	1962	1176	307	554	315
7	1169	pravi	1620	983	306	406	271
8	405	kljub	991	756	192	369	195
9	1592	težava	1317	754	192	349	213
10	1595	težek	1025	726	190	324	212
11	144	del	878	626	183	247	196
12	600	narediti	808	559	183	240	136
13	504	majhen	926	626	181	256	189
14	1308	proti	1655	915	179	490	246
15	1757	vrsta	891	680	177	278	225
16	1099	pomoč	1016	604	171	285	148
17	430	kratek	629	501	164	177	160
18	1447	slab	1100	683	148	346	189
19	828	nič	875	615	145	327	143
20	1743	volja	601	408	130	124	154
21	415	končati	624	467	126	188	153
22	1498	sprememba	830	471	122	215	134
23	1390	roka	625	452	119	207	126
24	114	cena	1116	461	110	172	179
25	879	očiten	570	445	108	241	96
26	1724	verjeten	549	422	108	211	103
27	1753	vmiti	465	354	108	157	89
28	535	moč	566	422	107	160	155
29	1463	smeti	473	379	102	196	81
30	1512	stati	506	377	99	161	117

Slika 5.3: Najpogostejših 30 negativnih besed slovarja Alfa.



Slika 5.4: Histogram najpogostejših 30 negativnih besed slovarja Alfa, ki se pojavijo v negativnih člankih.

5.5 Preveden slovar Beta

Po revizji je slovar pozitivnih besed štel 1669 besed, slovar negativnih pa 1912. Z izboljšanim slovarjem smo nato ponovno izvedli leksikalno metodo. V Tabeli 5.16 se nahajajo vrednosti klasifikacijske točnosti za klasifikator s slovarjem Beta, glede na izbrani mejni vrednosti. Najboljša vrednost je 0,46 pri vrednostih parametrov $m_p = 0,26$ in $m_n = 0$. Točnost je za 0,05 višja kot pri večinskem klasifikatorju. Z večanjem parametrov m_p in m_n se tudi v tem primeru točnost manjša.

mejne vrednosti	točnost
$m_p = m_n = 0$	0,43
$m_p = 0,26,\ m_n = 0$	0,46
$m_p = 0,10,\ m_n = 0,10$	0,43
$m_p = 0,25,\ m_n = 0,25$	0,41
$m_p = 0,50,\ m_n = 0,50$	0,35
$m_p = 0,75,\ m_n = 0,75$	0,32

Tabela 5.16: Klasifikacijske točnosti za različne vrednosti parametrov m_p in m_n za klasikator s slovarjem Beta.

	napovedani razred			
pravi razred	pozitiven	nevtralen	negativen	vsota
pozitiven	692	409	248	1349
nevtralen	618	558	427	1603
negativen	347	635	1066	2048
vsota	1657	1602	1741	5000

Tabela 5.17: Matrika zmot za klasifikator s slovarjem Beta, z mejnimi vrednostmi $m_p = 0,26$, $m_n = 0$.

mere	vrednost
točnost	0,46
MAE	0,66
MSE	0,89
mera $F1_{avg} = \frac{F1pozitiven + F1negativen}{2}$	0,51
T1 (ACC1)	0,88

Tabela 5.18: Rezultati mer za ocenjevanje klasifikatorja s slovarjem Beta.

5.6 Diskusija rezultatov

mere	večinski	originalni slovar	slovar Alfa	slovar Beta
točnost	0,41	0,43	0,35	**0,46**
MAE	0,86	0,70	0,86	**0,66**
MSE	1,4	0,98	1,28	**0,89**
mera $F1_{avg}$	0,29	0,48	0,29	**0,51**
T1 (ACC1)	0,73	0,86	0,79	**0,88**

Tabela 5.19: Primerjava mer za ocenjevanje za vsak klasifikator.

Kljub temu, da je po raziskavah [16] točnost leksikalnih metod nižja od točnosti metod strojnega učenja, so končni rezultati, ki so predstavljeni v Tabeli 5.19 pod pričakovanimi. Iz Tabele 5.19 lahko vidimo, da pri ocenjevanju prevedenih slovenskih novic z originalnim slovarjem dosežemo klasifikacijsko točnost 43%. To uporabnost angleškega slovarja dvigne nad večinski klasifikator, a le za 0,02. Vzporedno z večjo klasifikacijsko točnostjo se zvišata tudi mera $F1_{avg}$, ki ocenjuje priklic in preciznost na obeh pomembnih razredih (pozitivnem in negativnem), za 0,19 in pa T1 (ACC1) za 0,13. Tudi napaki MAE in MSE se v primerjavi z napakama večinskega klasifikatorja zmanjšata, a še vedno ostaneta na 0,70 in 0,98. Pri ocenjevanju slovenskih novic s prevedenim slovarjem Alfa, s poskušanjem različnih mejnih vrednosti dosežemo le 35% klasifikacijsko točnost. Točnost

večinskega klasifikatorja je v našem primeru 41%, kar pomeni, da je naš klasifikator s slovarjem Alfa neuporaben, saj je bolje uporabiti klasifikator, ki trdi da so vsi primeri negativni, kot pa uporabiti naš slovar Alfa in leksikalno metodo. Zaradi slabih rezultatov slovarja Alfa, smo se odločili, da slovar izboljšamo. Še enkrat smo temeljito pregledali slovarja in odstranili sporne besede. S pomočjo izboljšanega slovarja Beta smo nato ponovno izvedli analizo razpoloženja.

Pri vpeljavi izboljšav smo opazili, da so naša predvidevanja ob analizi prvih rezultatov napačna. Predvidevali smo, da je slovar pozitivnih besed bolje zgrajen kot slovar negativnih besed. V resnici je bil za porazne rezultate v veliki meri kriv slovar pozitivnih besed, saj je vseboval besedi *'za'* in *'dati'*, in je klasifikator s slovarjem Alfa posledično klasificiral večino člankov v razred pozitiven. Po vpeljavi izboljšav menimo, da je seznam negativnih besed bolje preveden kot seznam pozitivnih besed oziroma vsebuje več pomembnih negativnih besed za slovenski jezik, kot pa seznam pozitivnih besed pomembnih pozitivnih besed. Izboljšave so se izkazale za dobre saj je klasifikator s slovarjem Beta po primerjavi mer iz Tabele 5.19 v vseh merah najboljši. Z njim smo dosegli 46% klasifikacijsko točnost, vrednosti napak MAE - 0,66 in MSE - 0,89 sta v primerjavi z drugimi klasifikatorji manjši. Primerljivi sta z napakama klasifikatorja z originalnim angleškim slovarjem, in sta manjši, čeprav le za nekaj odstotkov. Skupaj s klasifikacijsko točnostjo se dvigne tudi vrednosti $F1_{avg}$ na 0,51. T1 (ACC1), ki doseže 0,88, nam pove, da bo naš klasifikator v 88% uvrstil naključen članek v pravi ali nevtralen razred. V 12% pa ga bo uvrstil v nasprotni (negativni $\Leftrightarrow$ pozitivni) razred.

Primerjali smo tehniko prevajanja slovarja in tehniko prevajanja celotnih besedil. Naši rezultati so pokazali, da so po izboljšanju slovenskega slovarja rezultati teh dveh tehnik primerljivi in le za malenkost je metoda prevajanja slovarja tudi boljša. Verjetno bi se dalo še z nadaljnjim dodajanjem, odstranjanjem in urejanjem predvsem seznama pozitivnih besed dobiti še boljše rezultate. Kljub pomanjkljivostim je ta slovar dobra iztočnica za nadaljnje delo. Čeprav smo z izboljšanjem slovarja izboljšali tudi mere

ocenjevanja kakovosti slovarja, rezultati niso dobri. Vzroke za to pripisujemo ročnemu ocenjevanju, kjer bi bolj točne rezultate dobili, če bi imeli več neodvisnih ocenjevalcev posameznih člankov. Največji problem pa predstavlja sama domena raziskave. V diplomskem delu smo izvajali analizo razpoloženja nad različnimi članki, ki so zajeti iz vsakodnevnih novic in blogov. Ugotavljali smo orientacijo subjektivnosti za vsako besedilo posebej. Kar pomeni, da smo ocenjevali orientacijo glede na vsako posamezno temo v posameznem članku. Besedila so bila različne dolžine, namenjene različnim bralcem, napisane iz različnih razlogov (obveščanje o novicah, vremenu, športu, izražanje mnenja trenutnih dogodkov, filmov, opisi avtomobilov, elektronskih naprav, itd.). Domena dnevnih novic je že sama po sebi težka za obdelavo analize razpoloženja, ker je tako obširna in navadno nima tako stroge omejitve znakov, kot na primer sporočilo Twitter, ki ima omejitev 140 znakov. Kot smo lahko opazili na pristopu, ki ga omenjamo v začetnem poglavju 1.2 za slovenski jezik, so za domeno izbrani zapisi na omrežju Twitter. Prav tako se znotraj zapisov omejijo na neko temo (šport, Tina Maze, politika, Janša) in gledajo razpoloženje v zapisih s tega zornega kota. Iz tega sklepamo, da so kratki, omejeni zapisi veliko primernejši za takšno obdelavo, ker z njimi ponavadi želimo sporočiti naše stališče, čustva, razpoloženja in še več, ker imamo na voljo omejeno število znakov za sporočanje, to naredimo bolj jasno in nazorno.

Poglavje 6

Sklepne ugotovitve

V diplomskem delu smo predstavili analizo razpoloženja in leksikalno metodo, pri kateri smo uporabili ročno izdelan slovar, najprej v slovenščini na slovenskih besedilih, nato pa še v angleščini na prevedenih angleških besedilih. V okviru dela sta nastala slovar slovenskih subjektivnih besed in ročno označen korpus slovenskih novic.

Glavni del diplome je opis izdelave slovarja in poteka analize razpoloženja, s katero smo slovar skušali izboljšati. Najprej je bilo potrebno prevesti slovar v slovenski jezik, nato ročno označiti korpus besedil. Sledilo je ocenjevanje besedil s pomočjo slovarjev in nato primerjava rezultatov z ročno označenim besedilom. Za primerjavo smo želeli preveriti še prevod korpusa besedil namesto slovarja, zato smo naredili analizo razpoloženja še z originalnim angleškim slovarjem na besedilih, prevedenih z orodjem Google Translate. Tudi te rezultate smo primerjali z ročno označenim korpusom, nato pa obe analizi, slovensko in angleško, primerjali med seboj. Ugotovili smo, da bi z izboljšanjem slovarja dobili boljše rezultate, zato smo izvedli nekaj dodatnih poskusov, ki so slovenski slovar do neke mere izboljšali. Rezultati so pokazali, da sta angleška in slovenska analiza primerljivi, slovenska pa se je izkazala malenkost bolje. Slovar in ročno označen korpus slovenskih novic sta dobri iztočnici za nadaljnje delo na področju analize razpoloženja v slovenskem jeziku, saj takšnih virov primanjkuje. Označen korpus, ki je nastal v okviru

dela, bi lahko razširili na več ocen na članek, če bi imeli na voljo več ocenjevalcev. Čeprav smo slovar do neke mere izboljšali že v tem delu, bi ga lahko izboljšali še s profesionalnim prevajanjem in pregledom jezikoslovcev. Naučili smo se, kako časovno zahtevna je izdelava dobrega slovarja, čeprav si pomagamo s prevajanjem obstoječega gradiva. Končen izboljšan slovar za slovenski jezik je dostopen na spletnem naslovu newstream.ijs.si/slosenti/.

Literatura

[1] M. Annett and G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie blogs," *Advances in artificial intelligence*, 2008.

[2] V. Beal. (2015) HTML - HyperText Markup Language. [Online]. Available: http://www.webopedia.com/TERM/H/HTML.html

[3] E-gradiva.net. (2015) Microsoft SQL Server. [Online]. Available: http://www.egradiva.net/drugo/omrezja/70_strezniki/02_datoteka.html

[4] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, pp. 84–89, apr 2013.

[5] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Advances in Artificial Intelligence*. Springer, 2009, pp. 207–210.

[6] T. P. group. (2015) What is PHP? [Online]. Available: http://php.net/manual/en/intro-whatis.php

[7] R. Hurwitz. (2013) Harvard General Inquirer. [Online]. Available: http://www.wjh.harvard.edu/~inquirer/homecat.htm

[8] M. Jurišič. (2013) Multilingual open source lemmatisation framework. [Online]. Available: http://lemmatise.ijs.si/

[9] M. Jurišič, I. Mozetič, T. Erjavec, and N. Lavrač, "LemmaGen : multilingual lemmatisation with induced Ripple-Down rules." *J. univers. comput. sci (Print) vol. 16, no. 9*, pp. 1190–1214, 2010.

[10] S. Kirithchenko, X. Zhu, and S. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, pp. 723–762, 2014.

[11] B. Škoda, "Rudarjenje razpoloženja na komentarjih rtvslo.si," Master's thesis, 2013.

[12] I. Kononenko and M. R. Šikonja, *Inteligentni sistemi.* Založba FE in FRI, 2010.

[13] B. Liu, *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, May 2012.

[14] B. Magnini, E. Pianta, L. Bentivogli, C. Girardi, and M. Speranza. (2015) WordNet Domains. [Online]. Available: http://wndomains.fbk.eu/

[15] R. Martinc, "Merjenje sentimenta na družbenem omrežju Twitter: izdelava orodja ter evalvacija," Master's thesis, 2013.

[16] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," *Proceedings of the 45th Annual Meeting of the Association of Computational Lunguistics, pages 976-983*, 2007.

[17] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *Big things come in small packages 718 in CEUR Workshop Proceedings.*, pp. 93–98, 2011.

[18] F. Å. Nielsen. (2015) AFINN. [Online]. Available: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[19] C. U. of Florida. (2015) ANEW. [Online]. Available: http://csea.phhp.ufl.edu/mission.html

[20] Z. SAZU and I. za slovenski jezik Frana Ramovša. (2015) Slovar slovenskega knjižnega jezika [elektronski vir]. [Online]. Available: http://bos.zrc-sazu.si/sskj.html

[21] SemEval. (2015) Semeval-2015 : Semantic evaluation exercises international workshop on semantic evaluation. [Online]. Available: http://alt.qcri.org/semeval2015/

[22] C. Strapparava and A. Valitutti. (2015) WordNet Affect. [Online]. Available: http://wndomains.fbk.eu/wnaffect.html

[23] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More Than Words: Quantifying Language to Measure Firms' Fundamentals ," *The Journal of Finance*, vol. 63, pp. 1437–1467, 2008.

[24] P. University. (2015) What is WordNet? [Online]. Available: http://wordnet.princeton.edu/

[25] M. Verlič, "Hibridni pristop za zaznavo elementov subjektivnosti v besedilnih tokovih," Ph.D. dissertation, 2009.

[26] W3Schools. (2015) HTML introduction. [Online]. Available: http://www.w3schools.com/html/html_intro.asp

[27] W3Schools. (2015) JavaScript tutorial. [Online]. Available: http://www.w3schools.com/js/js_whereto.asp

[28] W3Schools. (2015) PHP 5 tutorial. [Online]. Available: http://www.w3schools.com/php/default.asp

[29] Wikipedija. (2015) Besedilni korpus. [Online]. Available: http://sl.wikipedia.org/wiki/Besedilni_korpus

[30] Wikipedija. (2015) F1 score. [Online]. Available: http://en.wikipedia.org/wiki/F1_score

[31] Wikipedija. (2015) Hadamard product (matrices). [Online]. Available: http://en.wikipedia.org/wiki/Hadamard_product_(matrices)

[32] Wikipedija. (2015) Harmonična sredina. [Online]. Available: http://sl.wikipedia.org/wiki/Harmoni%C4%8Dna_sredina